Balancing Data Privacy and Utility: Benefits and Challenges in Developing a Synthetic Version of the Maryland SLDS

NCES STATS-DC
July 26th, 2019
Outline

● Overview of MLDSC

● Overview of Synthetic Data Project

● Evaluation of Research Utility

● Evaluation of Disclosure Risk
The MLDSC

- Independent state agency
- Receives, matches and merges education and workforce data from 3 partner state agencies: MSDE (grades K-12), MHEC (postsecondary), & DLLR (wages)
- Mission: Produce research reports and dashboards to inform state policy, programming, and the public
- Data are confidential, sensitive, and PII:
  - Data confidentiality protected by federal and state laws
  - Access to data granted to MLDSC staff only
The Synthetic Data Project

- 2015 SLDS grant from the U.S. Department of Education, Institute of Education Sciences ($2.7M) to create synthetic data version of the MLDS data. Aims:
  1. Create strategy to balance data confidentiality with need to make data available, and
  2. Expand access to the data to leverage research value.
- Synthetic data are generated based on models to mimic the relational patterns among variables
- Statistical analyses yield findings substantially similar to the real data, and
- simultaneously reduce the risk of privacy breach.
The Synthetic Data Project

1. Create gold standard datasets (GSDS)
   1. Study the data and potential uses
   2. Define GSDS
2. Synthesize GSDS
3. Evaluate research utility and disclosure risk

Next Steps:
1. Governing Board Approval
   1. Beta Testing – RU & DR
   2. Release synthetic data
2. Report on the project to inform other state longitudinal data systems
1. Gold Creation

1.1 Data Study
1.2 Evaluation of Existing Research Questions
1.3 End User input about research questions and methods
1.4 Definition of Cohort and Variables
1.5 Pass stakeholder review?

No

3. Evaluation

3.1 GSDS Utility Assessment
3.2 Synthetic Data Research Utility Assessment
3.3 Disclosure Risk Assessment

GSDS

2. Synthesization

SDS

3.4 Governing Board Approval
Creating the GSDS

Operational Data Store (ODS) (v=460)

Gold Standard Data Set (GSDS) (v=65, 50, 55)
(But there are many rows of data per person!)
10th grade GPA

1st quarter wages

12th grade GPA
Synthesization (Step 2)

- We need to satisfy a triangular trade-off:
  - Low (no) disclosure risk
  - Preservation of unconditional distributions
  - Preservation of multivariate conditional distributions
Synthesization

Gold Standard Data Set (GSDS) (v=65, 50, 55)  
Transformed (v=4000, 4700, 5900)

For synthesis, we need one wide record per individual

\[
a_1 a_2 a_3 a_4 a_5 a_6 c_1 c_2 c_3 c_4 c_5 c_6 D_1 D_2 \ldots \max(b)
\]
Synthesization

- Given the number of variables, potential interactions and non-linearities, and after initial testing and evaluation of existing methods, the decision was made to implement the CART method (Reiter, 2005b).

- CART is a method to model a dependent variable conditionally to a set of predictor variables.

- We have fully synthesized 3 versions of the data for our three GSDS.

- Final product will contain 30 synthesis datasets for each GSDS.

- We are currently evaluating the research utility and disclosure risk of the three versions of the 3 synthetic data sets.
1. Gold Creation

1.1 Data Study
1.2 Evaluation of Existing Research Questions
1.3 End User input about research questions and methods
1.4 Definition of Cohort and Variables
1.5 Pass stakeholder review?
   - No
   - Yes

2. Synthesization

3. Evaluation

3.1 GSDS Utility Assessment
3.2 Synthetic Data Research Utility Assessment
3.3 Disclosure Risk Assessment
3.4 Governing Board Approval

GSDS
SDS
Evaluation of synthetic data

- Synthetic data research utility assessment
  - Do you get the “right” answer from the synthetic data?

- Disclosure risk assessment
  - Do the synthetic data pose a risk of disclosure?
Scope of GSDS

- GSDS is comprised of data from:
  - High school students that entered the workforce
  - High school students that enrolled in post-secondary programs
  - Post-secondary students that entered the workforce

- In total, ~100 unique variables in the GSDS
  - Measures for many aspects of education in high school and post-secondary programs
  - Repeated measures for individuals on many variables over time (e.g., GPA, wages)
Utility Assessment

- Comparisons of variable distributions
  - Histograms and density plots
Utility Assessment

- Comparisons of variable distributions
  - Quantile plots

![SAT Math Comparison Diagram]

- The diagram shows quantile plots for SAT Math scores across different datasets (GSDS, SDS 1, SDS 2, SDS 3) compared to theoretical quantiles.

- The x-axis represents theoretical quantiles, while the y-axis represents sample quantiles.
Utility Assessment

- Comparisons of descriptive statistics
  - Means and standard deviations
  - Ranges for continuous, factor levels for categorical
  - Proportions of missing values
  - Correlations, contingency tables

- Evaluate within subgroups (e.g., Male/Female)
Utility Assessment - Specific

- How well does synthetic data reproduce the results of specific analyses?

- Gold standard analyses
  - Standardized mean differences
  - Bivariate correlations
  - Multiple regression
  - Logistic regression
  - Time series
Utility Assessment - Specific

- To illustrate components of specific utility assessment, we use a subset of the PS->WF GSDS and three SDSs.
- Regressed (log transformed) 2016 wages on gender, SAT-Math, transformed 2015 wages, and race/ethnicity categories
- The sample size of this cohort was 51,863 students
- We calculate the standardized difference between the estimates of interest based on the GSDS and for each SDS as

\[
SD = \frac{\beta_{SDS} - \beta_{GSDS}}{SE_{GSDS}}
\]
Utility Assessment - Specific

- We also calculate the measure of confidence interval overlap for each estimate (Karr, Kohnen, Organian, Reiter, & Sanil, 2006) as

\[
IO = 0.5 \left\{ \frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{GSDS} - LCL_{GSDS}} + \frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{SDS} - LCL_{SDS}} \right\}
\]

- where \( UCL_{SDS} \) and \( LCL_{SDS} \) represent, respectively, the average upper and lower confidence limits for the replicated estimates based on the SDSs and where \( UCL_{GSDS} \) and \( LCL_{GSDS} \) are the confidence limits for the estimate based on the GSDS.

- Note that when the two confidence intervals do not overlap, the further they are away from each other the more negative the \( IO \) estimate will become.
Utility Assessment - Specific

![Graph showing standardized coefficients for different variables with error bars.](Image)
## Utility Assessment - Specific

<table>
<thead>
<tr>
<th>Predictors</th>
<th>GSDS B (SE)</th>
<th>AVG SDS B (SE)</th>
<th>SD</th>
<th>IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable 1</td>
<td>0.446 (0.014)</td>
<td>0.343 (0.033)</td>
<td>7.572</td>
<td>-0.152</td>
</tr>
<tr>
<td>Variable 2</td>
<td>0.001 (0.012)</td>
<td>0.047 (0.014)</td>
<td>3.823</td>
<td>0.107</td>
</tr>
<tr>
<td>Variable 3</td>
<td>-0.065 (0.014)</td>
<td>-0.001 (0.018)</td>
<td>4.526</td>
<td>-0.018</td>
</tr>
<tr>
<td>Variable 4</td>
<td>-0.031 (0.012)</td>
<td>-0.007 (0.015)</td>
<td>1.912</td>
<td>0.568</td>
</tr>
<tr>
<td>Variable 5</td>
<td>0.001 (0.014)</td>
<td>-0.004 (0.015)</td>
<td>0.358</td>
<td>0.914</td>
</tr>
<tr>
<td>Variable 6</td>
<td>0.043 (0.014)</td>
<td>0.01 (0.016)</td>
<td>2.365</td>
<td>0.443</td>
</tr>
</tbody>
</table>
Utility Assessment - Specific

Wage Trajectories - HSWF

Actual Wages - Highest Wage Industry

Year

GSDS
Utility Assessment - Specific

- Cart model was not well tuned for wages
- Only one lag was used for employment in each sector
- Quarterly wage by sector was creating sparse data

- The solution that was implemented is the following:
  - All possible lags for wages are now used in the predictor set
  - Yearly global wage is synthesized first with all lags
  - Then quarterly percentages with all lags
  - Then sector percentage within quarterly with same sector lags and all quarters
Utility Assessment - Specific

Wage Trajectories - HSWF

GSDS

Actual Wages - Highest Wage Industry

Year

2012 2014 2016 2018
Utility Assessment - General

● How well does the synthetic data reproduce the variable relationships in the GSDS
  ○ Not tied to a specific analysis

● Several methods have been proposed
  ○ Kullback-Leibler divergence
  ○ Cluster analysis
  ○ Propensity scores
Utility Assessment - General

- Propensity score method

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Subj ID</th>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Variable 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>S1</td>
<td>1</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>S2</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>S3</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>S4</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Real Data

Synthetic Data
Utility Assessment - General

- Propensity score estimation
- Logistic regression
  - Interaction terms for higher-order moments
  - Generalized additive model
- Nonparametric classifier
  - CART
    - Naturally models interactions
  - Ensemble methods
    - Random forest
    - Boosted trees
Utility Assessment - General

- Overall measure of utility (Snoke, 2018; Woo, 2009)
  - Mean square error of propensity scores (pMSE)
    - $pMSE \rightarrow 0$, less discrepancy between real and synthetic datasets
  - Mostly used for comparing data synthesis methods

- Variable importance
  - Variables with high importance indicate discrepancies between the GSDS and SDS
Disclosure Risk Assessment

- Identification disclosure
  - relates to the potential for an intruder to match a given record with a specific individual

- Attribute disclosure
  - refers to the possibility that even aggregate data collected from these systems have the potential to disclose aspects of different subpopulations that may be sensitive in nature
Assessing Risk: Identification Disclosure

- Identification Disclosure rests on the assumption that the synthesized data contains identifiable information about individuals from the GSDS on which it was modeled.

- For fully synthesized data the “cases” do not exist (there are no “real” records), so theoretically, there is no risk of identity disclosure (the probability would conservatively be 1/N).

- One way to examine identification disclosure in fully synthesized data is to see if it is possible to determine if a specific record from the GSDS is in the SD.
# High School Cohort

<table>
<thead>
<tr>
<th>Category</th>
<th>Disclosure Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Disclosure Risk</td>
<td>0.000002</td>
</tr>
<tr>
<td>Disclosure Risk for Average Person (records near the median across categories)</td>
<td>0.000029</td>
</tr>
<tr>
<td>Known NAIC codes (NAIC=22 Utilities Sector)</td>
<td>0.001314</td>
</tr>
<tr>
<td>Population Uniques</td>
<td></td>
</tr>
<tr>
<td>- there are 824 instances where the array of characteristics are unique</td>
<td></td>
</tr>
<tr>
<td>- 284 of those have no counterpart in the synthetic data at all</td>
<td>0.209951</td>
</tr>
<tr>
<td>- 651 have no counterpart in at least one of the synthetic runs</td>
<td></td>
</tr>
</tbody>
</table>
Assessing Risk: Attribute Disclosure

- Attribute Disclosure relies on utilizing outside information (such as an additional dataset) to create inferences as a means to identify at-risk groups (<10)
- To assess the attribute disclosure risk we are using a subset of the original GSDS as our “outside source” of information
- The use of the original data provides a worst case scenario of external information an intruder might possess
- Disclosure risk is calculated as the odds of determining sensitive information (such as wages or test scores) using a process of probability matching between the synthetic and “outside” data
Disclosure Risk Assessment

- The below table examines the probability of identification of specific records in the synthesized data given specific levels of knowledge by an intruder. The information in the table is for demonstration purposes.

- The probabilities in the table were developed based on the methodology that is being utilized to calculate the disclosure risk for the synthetic data project but is based on simulations using 51,106 individual records from the Current Population Survey as described in a manuscript by Jerome P. Reiter (2005).

- The probabilities are calculated by dividing 1 over the total number of records identified as having the known characteristics.
Disclosure Risk Assessment

<table>
<thead>
<tr>
<th>Probabilities of Identification of a specific Record in Synthesized Data¹</th>
<th>Intruder knows...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic Characteristics</td>
<td>Demographic Characteristics and Educ outcomes</td>
</tr>
<tr>
<td>Intruder knows a specific record of interest is in the dataset².</td>
<td>0.00045</td>
</tr>
<tr>
<td>Intruder does not know a specific record of interest is in the dataset² and has knowledge of the underlying process used to synthesize data.</td>
<td>0.0016</td>
</tr>
</tbody>
</table>
Summary

- Public release of synthetic data has the potential to both create a safe and robust strategy to comply with data release requests, and, substantially expand access to the MLDS

- Research Utility
  - Multiple methods of assessment
  - Results inform data synthesis model

- Disclosure Risk
  - Identification and attribute disclosure
  - Because all variables are synthesized, in general disclosure risk is low
Next Steps

○ Request permission from MLDSC Governing Board for Beta Testing for Research Utility and Disclosure Risk

○ Incorporate lessons learned from RU and DR assessment and Beta testing into models

○ Seek permission to release synthetic data from MLDSC GB to release synthetic data

○ If given permission, build infrastructure and portal to create and maintain synthetic data and disseminate those data
Thank you!

● Contributors:
  ○ Daniel Bonnery, Yi Feng, Angie Henneberger, Tessa Johnson, Mark Lachowicz, Bess Rose, Terry Shaw, Laura Stapleton, Mike Woolley
  ○ Email: mlachowi@umd.edu
  mwoolley@ssw.umaryland.edu

● Acknowledgement:
  ○ This presentation was prepared by the Research Branch of the Maryland Longitudinal Data System Center (MLDSC) as part of funding from the U.S. Department of Education (R372A150045)
  ○ The Research Branch would like to thank the entire staff of the MLDSC for their assistance with the work and the presentation.
PLEASE COMPLETE THE EVALUATION!!

○ 4 ways:
  ○ Paper forms (submission boxes near registration table)
  ○ Go to https://www.surveymonkey.com/r/2019STATS-DC
  ○ Conference app under “Evaluation Forms”
  ○ Scan the QR code:

Session Number: 12-E